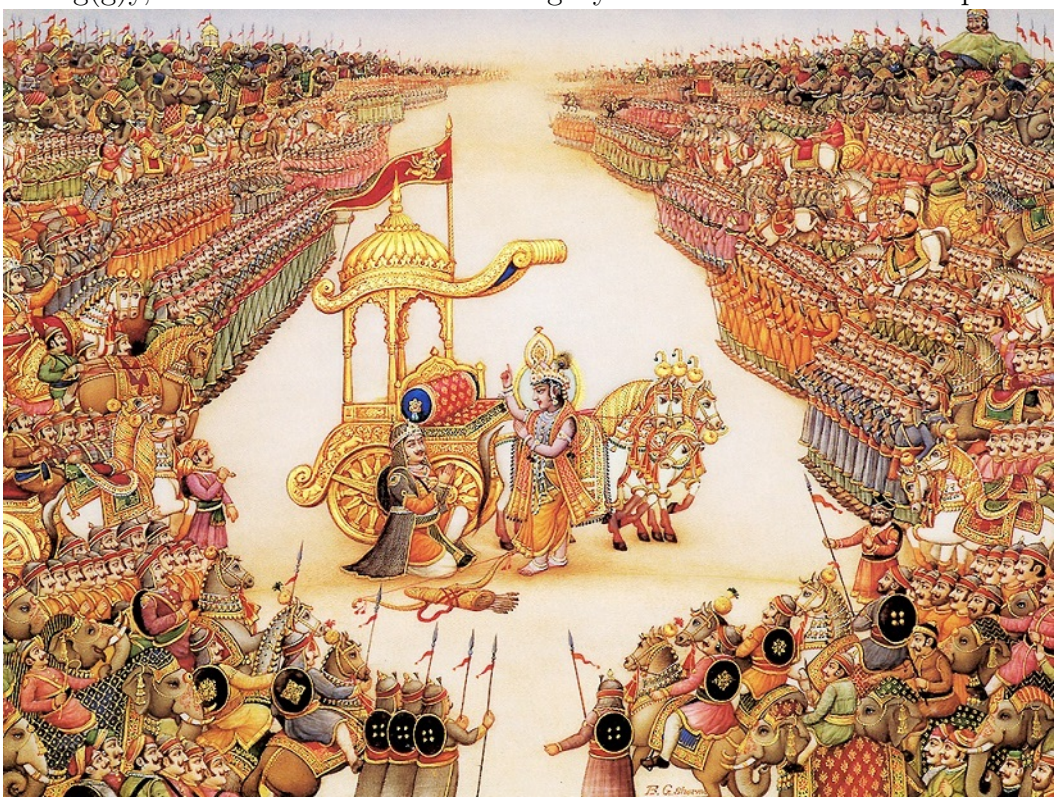
Pinter Consulting
New Series Nos. 25.

J K Pinter, Dr.Tech.

August 6, 2021

Motto

- Meg(g)y? Nem meg(g)y?
- Meg(g)y, de néha erőltetni kell az igényes matematikai továbbképzést.



- - - -

Introduction

Pinter Consulting of Calgary, Alberta practices Mathematics, promotes clear thinking and offers Consultations, Tutorials and Seminars in Mathematics.

Contents

21.0 Assignment 44.

Summary

- Mathematical Modelling
- *Ciarlet, Rózsa*
- Last revision August 6, 2021

Problem 1.

Consider the linear system whose matrix is block tridiagonal

$$\begin{bmatrix} B_1 & C_1 & & & \\ A_2 & B_2 & C_2 & & \\ & A_3 & B_3 & C_3 & \\ \dots & \dots & \dots & \dots & \dots \\ & & & A_N & B_N \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ \\ V_N \end{bmatrix} = \begin{bmatrix} D_1 \\ D_2 \\ \vdots \\ \\ D_N \end{bmatrix}. \tag{21.1}$$

With the provision of certain hypothesis, which should be stated explicitly, show that the solution of this linear system may be obtained by constructing successively the three sequences $\{Z\}, \{W\}, \{V\}$, the first of matrices, the other two of vectors,

$$Z_1 = B_1^{-1}C_1,\ Z_k = (B_k - A_k Z_{k-1})^{-1}C_k,\ k = 2, 3, \dots N,$$

$$W_1 = B_1^{-1}D_1,\ W_k = (B_k - A_k Z_{k-1})^{-1}(D_k - A_k W_{k-1}),\ k = 2, 3, \dots N,$$

$$V_N = W_N,\ V_k = W_k - Z_k V_{k+1}, k = N-1, N-2, \dots, 1$$

Proof:

$A_2, \dots A_N, B_1, B_2, \dots B_N$; and $C_1, C_2, \dots C_{N-1}$ are blocks of size $n \times n$ whereas $V_1, V_2, \dots V_N$, and $D_1, D_2, \dots D_N$ are vectors of size n. Note that $A's, B's, C's, D's$ are input.

The first block row is

$$B_1 V_1 + C_1 V_2 = D_1$$

Assuming that B^{-1} exists we have

$$V_1 + B_1^{-1}C_1V_2 = B_1^{-1}D_1$$

Write

$$Z_1 = B_1^{-1}C_1,\ W_1 = B_1^{-1}D_1,\ T_1^{-1} = B_1^{-1}.$$

Both matrix Z_1 and vector W_1 are computable, $\{T\}$ is an auxiliary set of matrices. Thus we are led to

$$V_1 + Z_1V_2 = W_1$$

and

$$V_1 = W_1 - Z_1V_2.$$

Next, we substitute this expression into second block row:

$$A_2V_1 + B_2V_2 + C_2V_3 = D_2$$

$$A_2(W_1 - Z_1V_2) + B_2V_2 + C_2V_3 = D_2$$

$$(B_2 - A_2Z_1)V_2 + C_2V_3 = D_2 - A_2W_1$$

Again, assuming the existence of the inverse of the auxiliary matrix T_2

$$T_2^{-1} = (B_2 - A_2Z_1)^{-1}$$

$$V_2 + (B_2 - A_2Z_1)^{-1}C_2V_3 = (B_2 - A_2Z_1)^{-1}(D_2 - A_2W_1)$$

Write

$$Z_2 = (B_2 - A_2Z_1)^{-1}C_2 = T_2^{-1}C_2$$

$$W_2 = (B_2 - A_2Z_1)^{-1}(D_2 - A_2W_1) = T_2^{-1}(D_2 - A_2W_1)$$

Then

$$V_2 + Z_2C_3 = W_2$$

and finally

$$V_2 = W_2 - Z_2V_3.$$

We can proceed to the next line

$$A_3V_2 + B_3V_3 + C_3V_4 = D_3$$

to obtain

$$V_3 = W_3 - Z_3V_4.$$

We repeat this procedure up to $N-1$. At each step

$$Z_k = (B_k - A_kZ_k)^{-1}C_k$$
$$W_k = (B_{N-}D_k - A_kW_{k-1})$$
$$V_k = W_{k-1} - Z_kV_{k+1}.$$
$$T_k^{-1} = (B_k - A_kZ_{k-1})^{-1}$$

where T_k is assumed to have an inverse. When we get to the last row

$$A_NV_{N-1} + B_NV_N = D_N$$

we can eliminate V_{N-1}

$$A_N(W_{N-1} - Z_{N-1}V_N) + B_NV_N = D_N$$
$$(B_N - A_NZ_{N-1})V_N = D_N - A_NW_{N-1}$$
$$V_N = (B_N - A_NZ_{N-1})^{-1}(D_N - A_NW_{N-1})$$
$$V_N = W_N$$

Thus V_N is resolved because W_N is computable at the $N-th$ step. Finally, backsubstitution into

$$V_k = W_{k-1} - Z_kV_{k+1}.$$

yields $V_{N-1}, V_{N-2}, \ldots V_2, V_1$, recursivelly. Therefore the solution to (1) can be calculated by constructing the following three sequences

$$Z_1 = B_1^{-1}C_1,\ Z_k = (B_k - A_kZ_{k-1})^{-1}C_k,\ k = 2, 3, \ldots N,$$
$$W_1 = B_1^{-1}D_1,\ W_k = (B_k - A_kZ_{k-1})^{-1}(D_k - A_kW_{k-1}),\ k = 2, 3, \ldots N,$$
$$V_N = W_N,\ V_k = W_k - Z_kV_{k+1}, k = N-1, N-2, \ldots, 1$$

provided that the fourth sequence exists:

$$T_1 = B_1^{-1},\ T_k = (B_k - A_kZ_{k-1})^{-1},\ k = 2, 3, \ldots N,$$

This is the equivalent of the standard algorithm for tridiagonal matrices. (cf. Thomas algorithm)

Demonstration:

Consider the following minimal 4×4 system:

$$\begin{bmatrix} b_1 & c_1 & & \\ a_2 & b_2 & c_2 & \\ & a_3 & b_3 & c_3 \\ & & b_4 & c_4 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \end{bmatrix}$$

First row:

$$b_1 v_1 + c_1 v_2 = d_1$$

$$v_1 = b_1^{-1}(d_1 - c_1 v_2),\ b_1 \neq 0$$

$$z_1 = b_1^{-1} c_1,\ w_1 = b_1^{-1} d_1,\ t_1^{-1} = b_1^{-1}.$$

$$v_1 + z_1 v_2 = w_1$$

$$v_1 = w_1 - z_1 v_2.$$

Second row:

$$a_2 v_1 + b_2 v_2 + c_2 v_3 = d_2$$

$$a_2(w_1 - z_1 v_2) + b_2 v_2 + c_2 v_3 = d_2$$

$$(b_2 - b_2 z_1) v_2 + c_2 v_3 = d_2 - a_2 w_1$$

$$t_2^{-1} = (b_2 - a_2 z_1)^{-1},\ (b_2 - a_2 z_1) \neq 0$$

$$v_2 + (b_2 - a_2 z_1)^{-1} c_2 v_3 = (b_2 - a_2 z_1)^{-1}(d_2 - a_2 w_1)$$

$$z_2 = (b_2 - a_2 z_1)^{-1} c_2 = t_2^{-1} c_2$$

$$w_2 = (b_2 - a_2 z_1)^{-1}(d_2 - a_2 w_1) = t_2^{-1}(d_2 - a_2 w_1)$$

$$v_2 + z_2 c_3 = w_2$$

$$v_2 = w_2 - z_2 v_3.$$

Third row:

$$a_3 v_2 + b_3 v_3 + c_3 v_4 = d_3$$

$$\vdots$$

$$v_3 = w_3 - z_3 v_4.$$

Fourth row:

$$a_4 v_3 + B_4 V_3 = d_4$$

$$a_4(w_3 - z_3 v_4) + b_4 v_4 = d_4$$

$$(b_4 - a_4 z_3) v_4 = d_4 - a_4 w_3$$

$$v_4 = (b_4 - a_4 z_3)^{-1}(d_4 - a_4 w_3)$$

$$v_4 = w_4 \surd$$

Recursion:

$$v_k = w_{k-1} - z_k v_{k+1}.$$

thus v_3, v_2, v_1 are computable if t_1, t_2, t_3, t_4 are not zero.

Problem 2.

Let A, B, I be $n \times n$ matrices. (I is the identity matrix.) Suppose $I - AB$ is invertible. Then

$$(I - BA)^{-1} = I + B(I - AB)^{-1}A.$$

Proof:

$$(I - BA)B = B - BAB = B(I - AB)$$

$$(I - BA)B(I - AB)^{-1}A = (B - BAB)(I - AB)^{-1}A =$$

$$B(I - AB)(I - AB)^{-1}A = BA$$

$$(I - BA) + (I - BA)B(I - AB)^{-1}A = (I - BA) + BA$$

$$(I - BA)(I + B(I - AB)^{-1}A) = I$$

$$(I - BA)(I + B(I - AB)^{-1}A) = I$$

$$(I - BA)^{-1} = I + B(I - AB)^{-1}A.$$

q.e.d.

Problem 3.

$$(A + uv^T)^{-1} = A^{-1} - \frac{(A^{-1}u)(v^T A^{-1})}{1 + v^T A^{-1} u}$$

Discussion:

This is the *Sherman - Morrison* formula. A is an invertible $n \times n$ matrix, u, v are compatible column vectors. Suppose A is modified by a diadic product and the new matrix is $(A+uv^T)$. The formula provides an inverse for the new matrix. $(A + uv^T)$ is an $n \times n$ matrix, $(A^{-1}u)$ is column vector, the product $(v^T A^{-1})$ is a row vector. $(A^{-1}u)(v^T A^{-1})$ is a diadic product, an $n \times n$ matrix. $(1 + v^T A^{-1}u)$ is a (non-zero) scalar, $(v^T A^{-1}u)$ is a scalar, too. We will use the fact that a product of compatible matrices is associative. In particular

$$(uv^T)(A^{-1}u)(v^T A^{-1}) = u(v^T A^{-1}u)(v^T A^{-1}) = (v^T A^{-1}u)u(v^T A^{-1}).$$

Proof:

$$(A + uv^T)\left(A^{-1} - \frac{(A^{-1}u)(v^T A^{-1})}{1 + v^T A^{-1}u}\right) =$$

$$AA^{-1} + uv^T A^{-1} - \frac{(AA^{-1}u)(v^T A^{-1})}{1 + v^T A^{-1}u} - \frac{u(v^T A^{-1}u)(v^T A^{-1})}{1 + v^T A^{-1}u} =$$

$$I + uv^T A^{-1} - \frac{(uv^T A^{-1})}{1 + v^T A^{-1}u} - \frac{(v^T A^{-1}u)(uv^T A^{-1})}{1 + v^T A^{-1}u} =$$

$$I + uv^T A^{-1} - (uv^T A^{-1})\left(\frac{I}{1 + v^T A^{-1}u} + \frac{(v^T A^{-1}u)I}{1 + v^T A^{-1}u}\right) =$$

$$I + (uv^T A^{-1})\left(I - \frac{I}{1 + v^T A^{-1}u} - \frac{(v^T A^{-1}u)I}{1 + v^T A^{-1}u}\right) =$$

$$I + (uv^T A^{-1})\left(\frac{1 + v^T A^{-1}u}{1 + v^T A^{-1}u}I - \frac{I}{1 + v^T A^{-1}u} - \frac{(v^T A^{-1}u)I}{1 + v^T A^{-1}u}\right) =$$

$$I + (uv^T A^{-1})\left(\frac{1 + v^T A^{-1}u}{1 + v^T A^{-1}u} - \frac{1}{1 + v^T A^{-1}u} - \frac{(v^T A^{-1}u)}{1 + v^T A^{-1}u}\right) I =$$

$$I + (uv^T A^{-1})\left(\frac{0}{1 + v^T A^{-1}u}\right) I = I.$$

Problem 4. Vector Norms in N-Dimensions

Definitions:

Write V for the *n-dimensional vector space* over the field $\mathbf{K}$ where $\mathbf{K}$ denotes the field of either real or complex numbers. A *norm* on V is a function

$$\| * \| : V \to \mathbf{R}$$

which satisfies the properties

$i)$ $\|v\| = 0 \iff v = 0,$ and $\|v\| \geq 0$ $\forall v \epsilon V$;

$ii)$ $\|\alpha v\| = |\alpha| \|v\|,$ $\alpha \epsilon \mathbf{K},$ $\mathrm{v} \epsilon \mathrm{V}$;

$iii)$ $\|u + v\| \leq \|u\| + \|v\|,$ $u, v \epsilon V$

Claim:

Let V be a finite-dimensional vector space. For every real number $p \geq 1$, the function $\| * \|$ defined by

$$\|v\|_p = \left(\sum_i |v_i|^p \right)^{\frac{1}{p}}$$

is a norm.

Discussion:

The non-negative property (i) and the homogeneous property (ii) of the norm is immediate, we shall prove them separately in other tutorial, here we deal with the triangle inequality (iii) for $p > 1$.

Lemma: p,q conjugates

Let $p > 1$ and set $q = p/(p-1)$. Note that

$$\frac{1}{p} + \frac{1}{q} = 1,$$

q is called the *conjugate* of p (and vice versa). From this equation we obtain a redundant set of equations

$$\frac{1}{p} = 1 - \frac{1}{q} = \frac{q-1}{q};$$

$$p = \frac{q}{q-1},\ p-1 = \frac{1}{q-1}.$$

Moreover

$$q = \frac{p}{p-1};\ q(p-1) = p;\ \frac{1}{q} = \frac{p-1}{p},$$

and finally

$$(p-1)q = p.$$

Lemma: Young's inequality

Let $p > 1$ and set $q = p/(p-1)$. For every $a, b \epsilon \mathbf{K}$,

$$|ab| \leq \frac{|a|^p}{p} + \frac{|b|^q}{q}.$$

Proof of Lemma

Since $(-\log)$ is a convex function then for every $\alpha,\ \beta > 0$

$$-\log\left(\frac{\alpha}{p} + \frac{\beta}{q}\right) \leq -\frac{1}{p}\log(\alpha) - \frac{1}{q}\log(\beta) = -\log(\alpha^{\frac{1}{p}}\beta^{\frac{1}{q}})$$

It follows that

$$\frac{\alpha}{p} + \frac{\beta}{q} \geq \alpha^{\frac{1}{p}}\beta^{\frac{1}{q}}.$$

Set $\alpha = |a|^p$, $\beta = |b|^q$ to recover the desired result. End of proof.$\surd$

Demonstration of Young's Lemma

Let $p > 1$, $q = \frac{p}{(p-1)}$, then

$$\frac{1}{p} + \frac{1}{q} = 1, \quad q - 1 = \frac{1}{p-1}$$

Consider the graph of function $f(x) = x^{p-1},\ p > 1$ for the positive quadrant of the co-ördinate system. This curve is convex,

$$y = x^{p-1}.$$

On this curve

$$x = y^{\frac{1}{(1-p)}} = y^{q-1}.$$

Now let x, y be two arbirtrary positive numbers, for example

$$y = OE,\ x = OD$$

Let us erect segments on the lines $AD : x = const$, and $EC : y = const$ up to their intersections with the curve, so we can see that the sum of the areas

of the figures OED a nd QAD is greater than the area of rectangle $OECD$, in other words

$$\int_0^x x^{p-1}dx + \int_0^y y^{q-1}dy \geq xy$$

or

$$\frac{x^p}{p} + \frac{y^q}{q} \geq xy.$$

The equality holds only for $x^p = y^q$. Moreover

$$(x-y)^2 \geq 0 \Rightarrow 2xy \leq x^2 + y^2$$

in its simplest form.

Lemma: Hölder's inequality

Let $p > 1$ and set $q = \dfrac{p}{p-1}$. For every x, y in V

$$\sum_{i=n} |x_i||y_i| \leq \|x\|_p \|y\|_q.$$

Proof:

Using the Young's inequality term-by-term:

$$\frac{\sum_{i=n} |x_i y_i|}{\|x\|_p \|y\|_q} = \sum_i \left(\frac{|x_i|}{\|x\|_p}\right)\left(\frac{|y_i|}{\|y\|_q}\right) \leq \frac{1}{p}\sum_{i=n}\left(\frac{|x_i|}{\|x\|_p}\right)^p + \frac{1}{q}\sum_{i=n}\left(\frac{|y_i|}{\|y\|_q}\right)^q$$

$$\sum_{i=n}\left(\frac{|x_i|}{\|x\|_p}\right)^p = 1, \quad \sum_{i=n}\left(\frac{|y_i|}{\|y\|_q}\right)^q = 1.$$

$$\frac{\sum_{i=n} |x_i y_i|}{\|x\|_p \|y\|_q} \leq \frac{1}{p} + \frac{1}{q} = 1$$

$$\sum_{i=n} |x_i y_i| \leq \|x\|_p \|y\|_q. \ \surd$$

Lemma: Minkowski's inequality

Let $p > 1$. For every x, y in V

$$\|x + y\|_p = \|x\|_p + \|y\|_p$$

Proof:

For every $i\epsilon\{1, 2, 3, \ldots n\}$, it follows from the triangle inequality

$$\begin{aligned} |x_i + y_i|^p &= |x_i + y_i||x_i + y_i|^{p-1} \\ &\leq |x_i||x_i + y_i|^{p-1} + |y_i||x_i + y_i|^{p-1} \end{aligned}$$

After summation we have

$$\sum_{i=1}^{n} |x_i + y_i|^p \leq \sum_{i=1}^{n} |x_i||x_i + y_i|^{p-1} + \sum_{i=1}^{n} |x_i||x_i + y_i|^{p-1}.$$

Next, we apply Hölder's inequality

$$\sum_{i=1}^{n} |x_i||x_i + y_i|^{p-1} \leq \|x_i\|_p * \||x_i + y_i|^{p-1}\|_q$$

$$\sum_{i=1}^{n} |y_i||x_i + y_i|^{p-1} \leq \|y_i\|_p * \||x_i + y_i|^{p-1}\|_q.$$

By definition of norm,

$$\||x_i + y_i|^{p-1}\|_q = \left(\sum_{i=1}^{n} ||x_i + y_i|^{p-1}|^q \right)^{\frac{1}{q}}.$$

Next, we manipulate the exponents by the **Lemma on conjugates.**

$$(p-1)q = p \text{ and } \frac{1}{q} = \frac{p-1}{p}.$$

Thus

$$\||x_i + y_i|^{p-1}\|_q = \left(\sum_i |x_i + y_i|^p \right)^{\frac{p-1}{p}} = \left(\left(\sum_i |x_i + y_i|^p \right)^{\frac{1}{p}} \right)^{p-1}.$$

$$\left(\sum_i |x_i + y_i|^p\right)^{\frac{1}{p}} = \||x_i + y_i\|_p$$

$$\||x_i + y_i|^{p-1}\|_q = (\|x_i + y_i\|_p)^{p-1}$$

This is an important partial result, let us substitute it into Hölder's inequality:

$$\sum_{i=1}^{n} |x_i||x_i + y_i|^{p-1} \leq \|x_i\|_p * (\||x_i + y_i\|_p)^{p-1}$$

$$\sum_{i=1}^{n} |y_i||x_i + y_i|^{p-1} \leq \|y_i\|_p * (\||x_i + y_i\|_p)^{p-1}$$

Adding up the x and y components yields

$$\sum_{i=1}^{n} |x_i + y_i|^p \leq (\|x\|_p + \|y\|_p) (\|x_i + y_i\|_p)^{p-1}.$$

Finally, we observe

$$\sum_{i=1}^{n} |x_i + y_i|^p = \left(\left(\sum_{i=1}^{n} |x_i + y_i|^p\right)^{\frac{1}{p}}\right)^p = (\||x_i + y_i\|_p)^p$$

hence

$$(\||x_i + y_i\|_p)^p \leq (\|x\|_p + \|y\|_p) (\|x_i + y_i\|_p)^{p-1}.$$

The last step is to cancel factor $(\|x_i + y_i\|_p)^{p-1}$ to arrive at the **Minkowski's inequality:**

$$\||x_i + y_i\|_p \leq \|x\|_p + \|y\|_p . \surd$$

Conclusion:

$$\|v\|_p = \left(\sum_i |v_i|^p\right)^{\frac{1}{p}}$$

is a norm.

21.1 Assignment 54.

Summary

- Mathematical Modelling and Numerical Analysis, Morton-Mayers
- Difference notation and truncation error
- Last revision August 6, 2021

Truncation error, example

$u_t = u_{xx}$

$0 = x_0 < x_1 < x_2 < \cdots < x_{J-1} < x_J = 1$

$\Delta x = x_{j+1} - x_j, \forall j, \text{equidistant}, \ \Delta t = t_{n+1} - t_n, \text{time step, fixed}$

Difference scheme:

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} = \frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{(\Delta x)^2}$$

Determine the principal part of the truncation error.

$$U_j^n \approx u(x_j, t_n),\ U_j^{n+1} \approx u(x_j, t_{n+1}),\ U_{j+1}^n \approx u(x_{j+1}, t_n),\ U_{j-1}^n \approx u(x_{j-1}, t_n)$$

U_j^n is the approximate value of the exact solution $u(x_j, t_n)$. Further, in the Taylor-expansion the derivaties are taken at (x_j, t_n).

$$U_j^{n+1} \approx u(x_j, t_n) + u_t \Delta t + \frac{1}{2} u_{tt} (\Delta t)^2 + \frac{1}{6} u_{ttt} (\Delta t)^3 + \frac{1}{24} u_{tttt} (\Delta t)^4 \ldots$$

$$U_{j+1}^n \approx u(x_j, t_n) + u_x \Delta x + \frac{1}{2} u_{xx} (\Delta x)^2 + \frac{1}{6} u_{xxx} (\Delta x)^3 + \frac{1}{24} u_{xxxx} (\Delta x)^4 \ldots$$

$$U_{j-1}^n \approx u(x_j, t_n) - u_x \Delta x + \frac{1}{2} u_{xx} (\Delta x)^2 - \frac{1}{6} u_{xxx} (\Delta x)^3 + \frac{1}{24} u_{xxxx} (\Delta x)^4 + \ldots$$

From the left-hand side of the difference scheme:

$$U_j^{n+1} - U_j^n = u_t \Delta t + \frac{1}{2} u_{tt} (\Delta t)^2 + \frac{1}{6} u_{ttt} (\Delta t)^3 + \frac{1}{12} u_{tttt} (\Delta t)^4 + \dots$$

From the other side

$$U_{j+1}^n - 2U_j^n + U_{j-1}^n =$$

$$u(x_j, t_n) + u_x \Delta x + \frac{1}{2} u_{xx} (\Delta x)^2 + \frac{1}{6} u_{xxx} (\Delta x)^3 + \frac{1}{24} u_{xxxx} (\Delta x)^4 + \dots$$

$$-2u(x_j, t_n)$$

$$+u(x_j, t_n) - u_x \Delta x + \frac{1}{2} u_{xx} (\Delta x)^2 - \frac{1}{6} u_{xxx} (\Delta x)^3 + \frac{1}{24} u_{xxxx} (\Delta x)^4 \cdots =$$

$$u_{xx} (\Delta x)^2 + \frac{1}{6} u_{xxxx} (\Delta x)^4 + \dots$$

Dividing the above expressions by Δt and Δx, respectively, we have

$$T(x,t) = (u_t - u_{xx}) + \left(\frac{1}{2} u_{tt} (\Delta t) - \frac{1}{12} u_{xxxx} (\Delta x)^2 \right) + \dots$$

Since $u(x, y)$ is the exact solution,

$$u_t - u_{xx} = 0,$$

and the principal part of the truncation error is

$$T(x,t) = \frac{1}{2} u_{tt} (\Delta t) - \frac{1}{12} u_{xxxx} (\Delta x)^2.$$

2.3

Suppose, that the mesh points x_j are chosen to satisfy

$$0 = x_0 < x_1 < x_2 < \cdots < x_{J-1} < x_J = 1$$

but otherwise arbitrary. The equation $u_t = u_{xx}$ is approximated over the interval $0 \le t \le t_F$ by

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} = \frac{2}{\Delta x_{j-1} + \Delta x_j} \left(\frac{U_{j+1}^n - U_j^n}{\Delta x_j} - \frac{U_j^n - U_{j-1}^n}{\Delta x_{j-1}} \right)$$

$$\Delta x_j = x_{j+1} - x_j.$$

Show that the leading terms of the truncation error of this approximations are

$$T_j^n = \frac{1}{2} \Delta t u_{tt} - \frac{1}{3} (\Delta x_j - \Delta x_{j-1}) u_{xxx} - \frac{1}{12} [(\Delta x_j)^2 + (\Delta x_{j-1})^2 - \Delta x_j \Delta x_{j-1}] u_{xxxx}.$$

Proof:

$$U_j^{n+1} - U_j^n = u_t \Delta t + \frac{1}{2} u_{tt} (\Delta t)^2 + \frac{1}{6} u_{ttt} (\Delta t)^3 + \frac{1}{24} u_{tttt} (\Delta t)^4 + \dots$$

$$U_{j+1}^n - U_j^n = u_x \Delta x_j + \frac{1}{2} u_{xx} (\Delta x_j)^2 + \frac{1}{6} u_{xxx} (\Delta x_j)^3 + \frac{1}{24} u_{xxxx} (\Delta x_j)^4$$

$$U_j^n - U_{j-1}^n = u_x \Delta x_{j-1} + \frac{1}{2} u_{xx} (\Delta x_{j-1})^2 + \frac{1}{6} u_{xxx} (\Delta x_{j-1})^3 + \frac{1}{24} u_{xxxx} (\Delta x_{j-1})^4$$

$$\frac{U_j^{n+1} - U_j^n}{dt} = u_t + \frac{1}{2} u_{tt} (\Delta t) + \frac{1}{6} u_{ttt} (\Delta t)^2 + \frac{1}{24} u_{tttt} (\Delta t)^3 + \dots$$

$$\frac{U_{j+1}^n - U_j^n}{\Delta x_j} = u_x + \frac{1}{2} u_{xx} (\Delta x_j) + \frac{1}{6} u_{xxx} (\Delta x_j)^2 + \frac{1}{24} u_{xxxx} (\Delta x_j)^3 + \dots$$

$$\frac{U_j^n - U_{j-1}^n}{\Delta x_{j-1}} = u_x - \frac{1}{2} u_{xx} (\Delta x_{j-1}) + \frac{1}{6} u_{xxx} (\Delta x_{j-1})^2 - \frac{1}{24} u_{xxxx} (\Delta x_{j-1})^3 + \dots$$

$$\frac{U_{j+1}^n - U_j^n}{\Delta x_j} - \frac{U_j^n - U_{j-1}^n}{\Delta x_{j-1}} =$$

$$\frac{1}{2} u_{xx} (\Delta x_j - \Delta x_{j-1}) + \frac{1}{6} u_{xxx} [(\Delta x_j)^2 - (\Delta x_{j-1})^2] + \frac{1}{24} u_{xxxx} [(\Delta x_j)^3 + (\Delta x_{j-1})^3]$$

$$[(\Delta x_j)^2 - (\Delta x_{j-1})^2] / [\Delta x_j + \Delta x_{j-1}] = \Delta x_j - \Delta x_{j-1}$$

$$[(\Delta x_j)^3 + (\Delta x_{j-1})^3] / [\Delta x_j + \Delta x_{j-1}] = (\Delta x_j)^2 - \Delta x_j \Delta x_{j-1} + (\Delta x_{j-1})^2$$

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} = u_t + \frac{1}{2} u_{tt} (\Delta t) + \frac{1}{6} u_{ttt} (\Delta t)^2 + \frac{1}{24} u_{tttt} (\Delta t)^3 + \dots =$$

$$u_{xx} + \frac{1}{3} u_{xxx} [(\Delta x_j) - (\Delta x_{j-1})] + \frac{1}{12} u_{xxxx} [(\Delta x_j)^2 - \Delta x_j \Delta x_{j-1} + (\Delta x_{j-1})^2] + \dots$$

Note that

$$u_{xx} = u_t.$$

$$T_j^n = u_t - u_{xx} + \frac{1}{3}u_{xxx}[(\Delta x_j) - (\Delta x_{j-1})] + \frac{1}{12}u_{xxxx}[(\Delta x_j)^2 - (\Delta x_j \Delta x_{j-1} + (\Delta x_{j-1})^2])$$
$$+ \frac{1}{2}u_{tt}(\Delta t) + \frac{1}{6}u_{ttt}(\Delta t)^2 + \frac{1}{24}u_{tttt}(\Delta t)^3 + \ldots$$

Principal part of truncation error includes terms up to u_{tt} and u_{xxxx}:

$$T_j^n = \frac{1}{2}u_{tt}(\Delta t) - \frac{1}{3}u_{xxx}[(\Delta x_j) - (\Delta x_{j-1})] - \frac{1}{12}u_{xxxx}[(\Delta x_j)^2 - (\Delta x_j \Delta x_{j-1} + (\Delta x_{j-1})^2]).$$

21.2 Assignment 56.

Summary

- *Pólya-Szegő : Aufgaben ...*
- Operation with power series (Henrici)
- Last revision August 6, 2021

Pólya-Szegő: Aufgaben.. Pt.I.No.34, Lemmas:

$$A) \sum_{k=0}^{\infty} a_k z^k \times \sum_{l=0}^{\infty} b_l z^l = \sum_{n=0}^{\infty} c_n z^n$$

where

$$c_n = a_0 b_n + a_1 b_{n-1} + a_2 b_{n-2} + \cdots + a_n b_0.$$

$$B) \sum_{k=0}^{\infty} \frac{\alpha_k}{k!} z^k \times \sum_{l=0}^{\infty} \frac{\beta_l}{l!} z^l = \sum_{n=0}^{\infty} \frac{\gamma_n}{n!} z^n$$

where

$$\gamma_n = \alpha_0 \beta_n + \binom{n}{1} \alpha_1 \beta_{n-1} + \binom{n}{2} \alpha_2 \beta_{n-2} + \cdots + \alpha_n \beta_0.$$

Problem:

Let $s_1, s_2, \ldots$ be complex numbers. Show that the reciprocal of the series

$$P := 1 - s_1 \frac{x}{1!} + \begin{vmatrix} s_1 & 1 \\ s_2 & s_1 \end{vmatrix} \frac{x^2}{2!} - \begin{vmatrix} s_1 & 1 & 0 \\ s_2 & s_1 & 2 \\ s_3 & s_2 & s_1 \end{vmatrix} \frac{x^3}{3!} + \ldots$$

is given by

$$Q := 1 + s_1 \frac{x}{1!} + \begin{vmatrix} s_1 & -1 \\ s_2 & s_1 \end{vmatrix} \frac{x^2}{2!} + \begin{vmatrix} s_1 & -1 & 0 \\ s_2 & s_1 & -2 \\ s_3 & s_2 & s_1 \end{vmatrix} \frac{x^3}{3!} + \ldots$$

[*Elemente der Math.*16, 87]

Discussion:

1st Claim:

$$\frac{Q'}{Q} = s_1 + s_2 x + s_3 x^2 + \dots$$

Write

$$\Delta_0 = 1;\ \Delta_1 = s_1;\ \Delta_2 = \begin{vmatrix} s_1 & -1 \\ s_2 & s_1 \end{vmatrix};\ \Delta_3 = \begin{vmatrix} s_1 & -1 & 0 \\ s_2 & s_1 & -2 \\ s_3 & s_2 & s_1 \end{vmatrix};\ \dots\ \Delta_n = \begin{vmatrix} s_1 & -1 & \dots & 0 \\ s_2 & s_1 & -2 & \dots \\ \dots & \dots & \dots & \dots \\ s_n & s_{n-1} & \dots & s_1 \end{vmatrix}$$

and

$$S = s_1 + s_2 x + s_3 x^2 + \dots$$

Then

$$Q = \Delta_0 + \frac{\Delta_1}{1!}x + \frac{\Delta_2}{2!}x^2 + \frac{\Delta_3}{3!}x^3 + \dots$$

and by term-by-term differentiation

$$Q' = \Delta_1 + \frac{\Delta_2}{1!}x + \frac{\Delta_3}{2!}x^2 + \frac{\Delta_4}{3!}x^3 + \dots$$

We wish to show

$$Q' = Q \times S = \left(\Delta_0 + \frac{\Delta_1}{1!}x + \frac{\Delta_2}{2!}x^2 + \frac{\Delta_3}{3!}x^3 + \dots\right) \times \left(s_1 + s_2 x + s_3 x^2 + \dots\right)$$

Coefficients of $Q \times S$, term-by-term

X^0**:**

$$\Delta_0 s_1 = s_1$$

X^1**:**

$$\Delta_0 1! s_2 \frac{x}{1!} + \Delta_1 \frac{x}{1!} s_1 = (s_2 + s_1 s_1)\frac{x}{1!} = \frac{1}{1!}\begin{vmatrix} s_1 & 1 \\ s_2 & s_1 \end{vmatrix} x$$

X^2:

$$\left(\Delta_0 s_3 + \frac{\Delta_1}{1!}s_2 + \frac{\Delta_2}{2!}s_1\right) = \frac{1}{2!}\left(2!\Delta_0 s_3 + 2!\frac{\Delta_1}{1!}s_2 + 2!\frac{\Delta_2}{2!}s_1\right) = \frac{1}{2!}\left(2!s_3 + 2s_1s_2 + s_1\Delta_2\right)$$

Notice that expansion by the last row gives

$$\begin{vmatrix} s_1 & -1 & 0 \\ s_2 & s_1 & -2 \\ s_3 & s_2 & s_1 \end{vmatrix} = s_3\begin{vmatrix} -1 & 0 \\ s_1 & -2 \end{vmatrix} - s_2\begin{vmatrix} s_1 & 0 \\ s_2 & -2 \end{vmatrix} + s_1\begin{vmatrix} s_1 & -1 \\ s_2 & s_1 \end{vmatrix} = s_3 2! + 2s_2s_1 + s_1(s_1^2 + ss_2),$$

which is equal to what we have calculated before:

$$\left(\Delta_0 s_3 + \frac{\Delta_1}{1!}s_2 + \frac{\Delta_2}{2!}s_1\right) = \frac{1}{2!}\begin{vmatrix} s_1 & -1 & 0 \\ s_2 & s_1 & -2 \\ s_3 & s_2 & s_1 \end{vmatrix}$$

X^3:

Expansion by last row:

$$\begin{vmatrix} s_1 & -1 & 0 & 0 \\ s_2 & s_1 & -2 & 0 \\ s_3 & s_2 & s_1 & -3 \\ s_4 & s_3 & s_2 & s_1 \end{vmatrix} = -s_4\begin{vmatrix} -1 & 0 & 0 \\ s_1 & -2 & 0 \\ s_2 & s_1 & -3 \end{vmatrix} + s_3\begin{vmatrix} s_1 & 0 & 0 \\ s_2 & -2 & 0 \\ s_3 & s_1 & -3 \end{vmatrix}$$

$$-s_2\begin{vmatrix} s_1 & -1 & 0 \\ s_2 & s_1 & 0 \\ s_3 & s_2 & -3 \end{vmatrix} + s_1\begin{vmatrix} s_1 & -1 & 0 \\ s_2 & s_1 & -2 \\ s_3 & s_2 & s_1 \end{vmatrix}.$$

Observations:

$$\begin{vmatrix} -1 & 0 & 0 \\ s_1 & -2 & 0 \\ s_2 & s_1 & -3 \end{vmatrix} = -3!\Delta_0$$

$$\begin{vmatrix} s_1 & 0 & 0 \\ s_2 & -2 & 0 \\ s_3 & s_1 & -3 \end{vmatrix} = 3!\Delta_1$$

$$\begin{vmatrix} s_1 & -1 & 0 \\ s_2 & s_1 & 0 \\ s_3 & s_2 & -3 \end{vmatrix} = (-3)\Delta_2$$

$$\begin{vmatrix} s_1 & -1 & 0 \\ s_2 & s_1 & -2 \\ s_3 & s_2 & s_1 \end{vmatrix} = \Delta_3$$

Therefore

$$\frac{1}{3!}\begin{vmatrix} s_1 & -1 & 0 & 0 \\ s_2 & s_1 & -2 & 0 \\ s_3 & s_2 & s_1 & -3 \\ s_4 & s_3 & s_2 & s_1 \end{vmatrix} = \frac{1}{3!}\left(s_4 3!\Delta_0 + s_3 3!\Delta_1 + s_2(-3)\Delta_2 + s_1\Delta_3\right)$$

or

$$\left(s_4\Delta_0 + s_3\Delta_1 + s_2\frac{\Delta_2}{2!} + s_1\frac{\Delta_3}{3!}\right) = \frac{1}{3!}\begin{vmatrix} s_1 & -1 & 0 & 0 \\ s_2 & s_1 & -2 & 0 \\ s_3 & s_2 & s_1 & -3 \\ s_4 & s_3 & s_2 & s_1 \end{vmatrix}.$$

X^4**:**

Expansion by the last row:

$$\begin{vmatrix} s_1 & -1 & 0 & 0 & 0 \\ s_2 & s_1 & -2 & 0 & 0 \\ s_3 & s_2 & s_1 & -3 & 0 \\ s_4 & s_3 & s_2 & s_1 & -4 \\ s_5 & s_4 & s_3 & s_2 & s_1 \end{vmatrix} = s_5 4! - s_4\begin{vmatrix} s_1 & 0 & 0 & 0 \\ s_2 & -2 & 0 & 0 \\ s_3 & s_1 & -3 & 0 \\ s_4 & s_2 & s_1 & -4 \end{vmatrix} + s_3\begin{vmatrix} s_1 & -1 & 0 & 0 \\ s_2 & s_1 & 0 & 0 \\ s_3 & s_2 & -3 & 0 \\ s_4 & s_3 & s_1 & -4 \end{vmatrix}$$

$$-s2\begin{vmatrix} s_1 & -1 & 0 & 0 \\ s_2 & s_1 & -2 & 0 \\ s_3 & s_2 & s_1 & 0 \\ s_4 & s_3 & s_2 & -4 \end{vmatrix} + s_1\Delta_4.$$

Observations:

$$\begin{vmatrix} s_1 & 0 & 0 & 0 \\ s_2 & -2 & 0 & 0 \\ s_3 & s_1 & -3 & 0 \\ s_4 & s_2 & s_1 & -4 \end{vmatrix} = -4!s_1 = -4!\Delta_1$$

$$\begin{vmatrix} s_1 & -1 & 0 & 0 \\ s_2 & s_1 & 0 & 0 \\ s_3 & s_2 & -3 & 0 \\ s_4 & s_3 & s_1 & -4 \end{vmatrix} = \begin{vmatrix} s_1 & -1 \\ s_2 & s_1 \end{vmatrix} * \begin{vmatrix} -3 & 0 \\ s_1 & -4 \end{vmatrix} = 3 * 4 * \Delta_2$$

$$\begin{vmatrix} s_1 & -1 & 0 & 0 \\ s_2 & s_1 & -2 & 0 \\ s_3 & s_2 & s_1 & 0 \\ s_4 & s_3 & s_2 & -4 \end{vmatrix} = -(-4)\Delta_3 = 4\Delta_3$$

$$\left(s_5\Delta_0 + s_4\frac{\Delta_1}{1!} + s_3\frac{\Delta_2}{2!} + s_2\frac{\Delta_3}{3!} + s_1\frac{\Delta_4}{4!} \right) = \frac{1}{4!}\Delta_5$$

X^5**:**

Expansion by the last row;

List of minors:

$n = 6$**;** $k = 1$ **:**

$$M_{6,1} = \begin{vmatrix} . & -1 & 0 & 0 & 0 & 0 \\ . & s_1 & -2 & 0 & 0 & 0 \\ . & s_2 & s_1 & -3 & 0 & 0 \\ . & s_3 & s_2 & s_1 & -4 & 0 \\ . & s_4 & s_3 & s_2 & s_1 & -5 \\ . & . & . & . & . & . \end{vmatrix} = \begin{vmatrix} -1 & 0 & 0 & 0 & 0 \\ s_1 & -2 & 0 & 0 & 0 \\ s_2 & s_1 & -3 & 0 & 0 \\ s_3 & s_2 & s_1 & -4 & 0 \\ s_4 & s_3 & s_2 & s_1 & -5 \end{vmatrix} = -5! = -(n-1)!\Delta_0$$

determinant of lower-triangular matrix

$n = 6$**;** $k = 2$**:**

$$M_{6,2} = \begin{vmatrix} s_1 & . & 0 & 0 & 0 & 0 \\ s_2 & . & -2 & 0 & 0 & 0 \\ s_3 & . & s_1 & -3 & 0 & 0 \\ s_4 & . & s_2 & s_1 & -4 & 0 \\ s_5 & . & s_3 & s_2 & s_1 & -5 \\ . & . & . & . & . & . \end{vmatrix} = \begin{vmatrix} s_1 & 0 & 0 & 0 & 0 \\ s_2 & -2 & 0 & 0 & 0 \\ s_3 & s_1 & -3 & 0 & 0 \\ s_4 & s_2 & s_1 & -4 & 0 \\ s_5 & s_3 & s_2 & s_1 & -5 \end{vmatrix} = (n-1)!\Delta_1$$

determinant of lower-triangular matrix , same structure as $M_{5,1}$

$n = 6$; $k = 3$:

$$M_{6,3} = \begin{vmatrix} s_1 & -1 & . & 0 & 0 & 0 \\ s_2 & s_1 & . & 0 & 0 & 0 \\ s_3 & s_2 & . & -3 & 0 & 0 \\ s_4 & s_3 & . & s_1 & -4 & 0 \\ s_5 & s_4 & . & s_2 & s_1 & -5 \\ . & . & . & . & . & . \end{vmatrix} = \begin{vmatrix} s_1 & -1 & 0 & 0 & 0 \\ s_2 & s_1 & 0 & 0 & 0 \\ s_3 & s_2 & -3 & 0 & 0 \\ s_4 & s_3 & s_1 & -4 & 0 \\ s_5 & s_4 & s_2 & s_1 & -5 \end{vmatrix} = \begin{vmatrix} s_1 & -1 \\ s_2 & s_1 \end{vmatrix} \begin{vmatrix} -3 & 0 & 0 \\ s_1 & -4 & 0 \\ s_2 & s_1 & -5 \end{vmatrix}$$

$$M_{6,3} = (-3)(-4)(-5)\Delta_2$$

Laplace theorem; principal minor of order 2 multiplied by complementary minor of triangular submatrix.

$n = 6$; $k = 4$:

$$M_{6,4} = \begin{vmatrix} s_1 & -1 & 0 & . & 0 & 0 \\ s_2 & s_1 & -2 & . & 0 & 0 \\ s_3 & s_2 & s_1 & . & 0 & 0 \\ s_4 & s_3 & s_2 & . & -4 & 0 \\ s_5 & s_4 & s_3 & . & s_1 & -5 \\ . & . & . & . & . & . \end{vmatrix} = \begin{vmatrix} s_1 & -1 & 0 & 0 & 0 \\ s_2 & s_1 & -2 & 0 & 0 \\ s_3 & s_2 & s_1 & 0 & 0 \\ s_4 & s_3 & s_2 & -4 & 0 \\ s_5 & s_4 & s_3 & s_1 & -5 \end{vmatrix} = \begin{vmatrix} s_1 & -1 & 0 \\ s_2 & s_1 & -2 \\ s_3 & s_2 & s_1 \end{vmatrix} \begin{vmatrix} -4 & 0 \\ s_1 & -5 \end{vmatrix}$$

$$M_{6,4} = (-4)(-5)\Delta_3$$

Laplace theorem; principal minor of order 3 multiplied by complementary minor of triangular submatrix.

$n = 6$; $k = 6$:

$$M_{5,2} = \begin{vmatrix} s_1 & -1 & 0 & 0 & 0 & . \\ s_2 & s_1 & -2 & 0 & 0 & . \\ s_3 & s_2 & s_1 & -3 & 0 & . \\ s_4 & s_3 & s_2 & s_1 & -4 & . \\ s_5 & s_4 & s_3 & s_2 & s_1 & . \\ . & . & . & . & . & . \end{vmatrix} = \begin{vmatrix} s_1 & -1 & 0 & 0 & 0 \\ s_2 & s_1 & -2 & 0 & 0 \\ s_3 & s_2 & s_1 & -3 & 0 \\ s_4 & s_3 & s_2 & s_1 & -4 \\ s_5 & s_4 & s_3 & s_2 & s_1 \end{vmatrix} = 4\Delta_5$$

X^n:

$$\left(s_n\Delta_0 + s_{n-1}\frac{\Delta_1}{1!} + s_{n-2}\frac{\Delta_2}{2!} \cdots + s_2\frac{\Delta_{n-2}}{(n-2)!} + s_1\frac{\Delta_{n-1}}{(n-1)!} \right) = \frac{\Delta_n}{(n-1)!}$$

Proof:

Expansion by the last row. Then apply Laplace theorem on first row; first and second row; first, second and third row, etc.

Numerical check:

Calculate

$$\left(s_5\Delta_0 + s_4\frac{\Delta_1}{1!} + s_3\frac{\Delta_2}{2!} + s_2\frac{\Delta_3}{3!} + s_1\frac{\Delta_4}{4!}\right) = \frac{1}{4!}\Delta_5$$

```
Henrici Tests
Factorials
1.000000   2.000000   6.000000  24.000000 120.000000
Input s
-0.848154   0.230834  -0.089694   0.775749  -0.532372
Input matrix
-0.848154  -1.000000   0.000000   0.000000   0.000000
 0.230834  -0.848154  -2.000000   0.000000   0.000000
-0.089694   0.230834  -0.848154  -3.000000   0.000000
 0.775749  -0.089694   0.230834  -0.848154  -4.000000
-0.532372   0.775749  -0.089694   0.230834  -0.848154
order=  0  delta=   1.000000
order=  1  delta=  -0.848154
order=  2  delta=   0.950200
order=  3  delta=  -1.376869
order=  4  delta=   6.936751
order=  5  delta= -36.745308
accu=  -1.531054
accu=  -1.531054
```

First line, factorials, second line input numbers, random numbers between -1.0 and 1.0; followed by the 5×5 input matrix. Then Δ_0, Δ_1 ... Δ_5 calculated by simple Gaussian elimination. Finally, *accu* denotes the right and left side of the equation. Equality demonstrated.

$$\text{accu} = \left(s_5\Delta_0 + s_4\frac{\Delta_1}{1!} + s_3\frac{\Delta_2}{2!} + s_2\frac{\Delta_3}{3!} + s_1\frac{\Delta_4}{4!}\right)$$

$$\text{accu} = \frac{\Delta_5}{4!}.$$

End of 1st Claim.

2nd Claim:

$$Q^{-1} = \exp\left(-\int_0^t S(\tau)d\tau\right)$$

Proof:

As shown before

$$\frac{Q'(t)}{Q(t)} = S(t).$$

Separable ordinary differential equation, solution:

$$\ln Q(t) = \int_0^t S(\tau)d\tau.$$

$$Q(t) = \exp\int_0^t S(\tau)d\tau$$

Thus the reciprocal of Q

$$\frac{1}{Q} = \exp\left(-\int_0^t S(\tau)d\tau\right).$$

2nd Claim proven.

3rd Claim:

$$Q^{-1} = P.$$

Proof:

Write $-s_i$ for s_i in $\Delta_n, n = 1, 2\ldots$ as shown by formula above.

$$Q^{-1} := 1 + (-s_1)\frac{x}{1!} + \begin{vmatrix} -s_1 & -1 \\ -s_2 & -s_1 \end{vmatrix}\frac{x^2}{2!} + \begin{vmatrix} -s_1 & -1 & 0 \\ -s_2 & -s_1 & -2 \\ -s_3 & -s_2 & -s_1 \end{vmatrix}\frac{x^3}{3!} + \ldots$$

$$(-s_1)\frac{x}{1!} \Rightarrow -s_1\frac{x}{1!}$$

$n = 2$

$$\begin{vmatrix} -s_1 & -1 \\ -s_2 & -s_1 \end{vmatrix} \left| \frac{x^2}{2!} \right. \Rightarrow (-1)(-1) \begin{vmatrix} s_1 & 1 \\ s_2 & s_1 \end{vmatrix} \left| \frac{x^2}{2!} \right.$$

$n = 3$

$$\begin{vmatrix} -s_1 & -1 & 0 \\ -s_2 & -s_1 & -2 \\ -s_3 & -s_2 & -s_1 \end{vmatrix} \left| \frac{x^3}{3!} \right. \Rightarrow (-1)^3 \begin{vmatrix} s_1 & 1 & 0 \\ s_2 & s_1 & 2 \\ s_3 & s_2 & s_1 \end{vmatrix} \left| \frac{x^3}{3!} \right.$$

These are the coefficients of $\frac{x^1}{1!}, \frac{x^2}{2!}$ and $\frac{x^3}{3!}$ in the formal power series P. All coefficients are determined in the same manner, notice the alternating sign for odd and even powers of x. This proves

$$Q^{-1} = P.$$

Appendix:

```
program elem
!
!   Numerical verification of
!   Elemente der Math.  16, 87
!
integer:: i,j,k,n
real:: s(5), r(5), x, accu
real:: delta(0:5), rfact(0:5)
real, parameter :: pi=3.1415927
real, dimension (:,:), allocatable :: darray
! dynamic array
real, dimension (1:5,1:5) :: p
!
open(unit=10,file="Henrici.dat",status="old",action="write",iostat=ierror)
if(ierror/=0) then
print*, "failed to open Henrici.dat"
stop
else
print*, "  ***   opened Henrci.dat"
```

```
end if
write(10,100)
100                 format(' Henrici Tests' )
!
s=0.0; delta=0.0; rfact=1.0
!
do i=1,5
call random_number(s)
end do
!
do i=1,5
call random_number(r)
end do
!
do i=1,5
s(i)=s(i)*sin(2.0*pi*r(i))
end do
!
do i=1,5
rfact(i)= rfact(i-1)*i
end do
write ( 10,  * ) " Factorials"
write ( 10, "( 5 f11.6 )" )  (rfact(j), j=1,5)
!
write ( 10,  * ) " Input s"
write ( 10, "( 5 f11.6 )" )  s
!
do i=1,5
do j=1,5
if(i-j.eq.4) p(i,j)=s(5)
if(i-j.eq.3) p(i,j)=s(4)
if(i-j.eq.2) p(i,j)=s(3)
if(i-j.eq.1) p(i,j)=s(2)
if(i-j.eq.0) p(i,j)=s(1)
if(i-j.lt.0) p(i,j)=0.0
end do
end do
!
```

```
p(1,2)=-1.0; p(2,3)=-2.0; p(3,4)=-3.0; p(4,5)=-4.0;
!
write ( 10,  * ) " Input matrix"
do i=1,5
write ( 10, "( 5 f11.6 )" )  (p(i,j), j=1,5)
end do
!
n=0.0
x=1.0
delta(n)=x
write ( 10, 1000 ) n,  x
!
do k=1,5
n=k
allocate(darray(n,n))
do i=1,n
do j=1,n
darray(i,j)=p(i,j)
end do
end do
!
call delta_1(darray,n,x)
delta(n)=x
write ( 10, 1000 ) n,  x
1000 format('  order=',i3,'  delta=',  f11.6 )
!
deallocate (darray)
end do
! calculate formula
!
accu=s(5)*delta(0)
do i=1,4
accu=accu+delta(i)*s(5-i)/rfact(i)
end do
!
write ( 10, 1001 ) accu
1001 format(' accu=',  f11.6 )
!
```

```
accu= delta(5)/rfact(4)
write ( 10, 1001 ) accu
end program elem
!
subroutine delta_1(a,n,x)
!=============================================================
! determinant, basic elimination
!-------------------------------------------------------------
! input ...
! a(n,n) - array of coefficients for matrix A
! n      - number of equations
! output ...
! x   - solution, evaluated determinant
! comments ...
! the original arrays a(n,n) will be destroyed
! during the calculation
!=============================================================
implicit none
integer n
real a(n,n)
real c,x
integer i, j, k

!step 1: forward elimination
do k=1, n-1
!
print*, "sweep", k
if(a(k,k).eq.0.0) then
print*, "zero in diag", k
stop
endif
!
do i=k+1,n

c=a(i,k)/a(k,k)
a(i,k) = 0.0
do j=k+1,n
a(i,j) = a(i,j)-c*a(k,j)
```

```
end do
end do
!! check array

!     do i = 1, n
!      write(*,*) a (i,1), a(i,2), a(i,3), a(i,4), a(i,5)
!     end do
!
end do

!step 2: multiplpy elements in the main diagonal
x=1.0d00
do i=1,n
x=x*a(i,i)
end do
end subroutine delta_1
!
subroutine  delta_2(a,n,x)
!============================================================
! Evaluate determinant
! Method: Gauss elimination with scaling and pivoting
!------------------------------------------------------------
! input ...
! a(n,n) - array of coefficients for matrix A
! n      - size of matrix A
! output ...
! x      - determinant
! coments ...
! the original arrays a(n,n) and b(n) will be destroyed
! during the calculation
!============================================================
implicit none
integer n
real, dimension (n,n) :: a
real, dimension (1:n) :: s
real c, pivot, store, x
integer i, j, k, l
! step 1: begin forward elimination
```

```
do k=1, n-1

! step 2: "scaling"
! s(i) will have the largest element from row i
do i=k,n                        ! loop over rows
s(i) = 0.0
do j=k,n                     ! loop over elements of row i
s(i) = max(s(i),abs(a(i,j)))
end do
end do

! step 3: "pivoting 1"
! find a row with the largest pivoting element
pivot = abs(a(k,k)/s(k))
l = k
do j=k+1,n
if(abs(a(j,k)/s(j)) > pivot) then
pivot = abs(a(j,k)/s(j))
l = j
end if
end do

! Check if the system has a sigular matrix
if(pivot == 0.0) then
write(*,*) ' The matrix is sigular '
return
end if

! step 4: "pivoting 2" interchange rows k and l (if needed)
if (l /= k) then
do j=k,n
store = a(k,j)
a(k,j) = a(l,j)
a(l,j) = store
end do

end if
```

```
! step 5: the elimination (after scaling and pivoting)
do i=k+1,n
c=a(i,k)/a(k,k)
a(i,k) = 0.0

do j=k+1,n
a(i,j) = a(i,j)-c*a(k,j)
end do
end do
end do

! step 6: determinant
x=1.0
do i=1,n
print*, a(i,i)
x=x*a(i,i)
end do
!
end subroutine delta_2
```

21.3 Assignment 57.

Summary

- *Pólya-Szegő : Aufgaben ...*
- Operation with power series (Henrici)
- Last revision August 6, 2021

Closed formula for Hessenberg determinant

$$\Delta_n = \begin{vmatrix} a_{11} & a_{12} & 0 & \dots & 0 \\ a_{21} & a_{22} & a_{23} & \dots & 0 \\ a_{31} & a_{32} & a_{33} & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & 0 \\ a_{n-1,1} & a_{n-1,2} & a_{n-1,3} & \dots & a_{n-1,n} \\ a_{n,1} & a_{n,2} & a_{n,3} & \dots & a_{n,n} \end{vmatrix}$$

$$\Delta_0 = 1$$

$$\Delta_1 = a_{11}$$

$$\Delta_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}$$

$$\Delta_n = a_{n,n}\Delta_{n-1} + \sum_{r=1}^{n-1} \left[(-1)^{n-r} a_{n,r} \Pi_{j=r}^{n-1} a_{j,j+1} * \Delta_{r-1}) \right]$$

Demonstration

Expansion by the last row, leading principal minor in the NW corner multiplied by apparent diagonal.

$$\Delta_n = \begin{vmatrix} a_{11} & a_{12} & 0 & \dots & \dots & \dots & 0 \\ a_{21} & a_{22} & a_{23} & 0 & \dots & \dots & 0 \\ a_{31} & a_{32} & a_{33} & a_{34} & \dots & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \dots & \dots & 0 \\ a_{n-1,1} & a_{n-1,2} & a_{n-1,3} & \dots & \dots & \dots & a_{n-1,n} \\ a_{n,1} & a_{n,2} & a_{n,3} & \dots & \dots & \dots & a_{n,n} \end{vmatrix}$$

Example: $r = 3$

$$\begin{vmatrix} a_{11} & a_{12} & * & 0 & \dots & \dots & 0 \\ a_{21} & a_{22} & * & 0 & \dots & \dots & 0 \\ a_{31} & a_{32} & * & a_{34} & \dots & \dots & 0 \\ \vdots & \vdots & * & \dots & \dots & \dots & 0 \\ a_{n-1,1} & a_{n-1,2} & * & \dots & \dots & \dots & a_{n-1,n} \\ * & * & * & * & \dots & \dots & * \end{vmatrix}$$

Laplace expansion by the first $(r-1)$ row yields

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = \Delta_2$$

Opposite to NW corner is apparent diagonal

$$(a_{34} * a_{4,5} * \cdots * a_{n-1,n})$$

Term: $r = 3$

$$\left[(-1)^{n-3} a_{n,3} \left(\Pi_{j=3}^{n-1} a_{j,j+1}\right) * \Delta_2)\right] \surd$$

Closed formula for Toeplitz-Hessenberg determinant

Lemma: Wronski-formula

If

$$P = a_0 + a_1 x + a_2 x^2 + \dots$$

is a formal power series with $a_0 \neq 0$; then the coefficients of the reciprocal series

$$P^{-1} = b_0 + b_1 x + b_2 x^2 + \dots$$

are given by

$$b_0 = \frac{1}{a_0}, \quad b_n = \frac{(-1)^n}{a_0^{n+1}} \det \begin{bmatrix} a_1 & a_0 & 0 & \dots & 0 \\ a_2 & a_1 & a_0 & \dots & 0 \\ a_3 & a_2 & a_1 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & 0 \\ a_{n-1} & a_{n-2} & \dots & a_1 & a_0 \\ a_n & a_{n-1} & a_{n-2} & \dots & a_1 \end{bmatrix}^T, \quad n \geq 1.$$

Proof:

$$PP^{-1} = 1$$

$$PP^{-1} = \sum_{n\geq 1} \left(\sum_{k=0}^{n} a_k b_{n-k} \right) x^n$$

$$\sum_{k=0} a_k b_{n-k} = \delta_{0,n}, \quad \text{(Kronecker-delta)}$$

or

$$\begin{bmatrix} a_0 & 0 & \dots & 0 \\ a_1 & a_0 & \dots & 0 \\ a_2 & \dots & \dots & 0 \\ \dots & \dots & a_1 & a_0 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \dots \\ b_n \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \dots \\ 0 \end{bmatrix}.$$

then b_k is calculated by Cramer's rule.

Lemma: Closed formula for b_k

Let P, R^{-1} be as before, then

$$\frac{b_n}{b_0} = \sum_T \binom{t_1 + t_2 + \dots t_n}{t_1, t_2, \dots t_n} \left(-\frac{a_1}{a_0}\right)^{t_1} \left(-\frac{a_2}{a_0}\right)^{t_2} \dots \left(-\frac{a_n}{a_0}\right)^{t_n},$$

where

$$T := 1 * t_1 + 2 * t_2 + +3 * t_3 + \dots = n, \text{ partition of n.}$$

Proof:

$$\frac{P^{-1}}{b_0} = \frac{1}{\frac{P}{a_0}} = \frac{1}{1 - \sum_{n>0} \left(-\frac{a_n}{a_0}\right) x^n} = 1 + \sum_{j>0} \left(\sum_{n>0} \left(-\frac{a_n}{a_0}\right) x^n \right)^j$$

Compare to

$$\frac{1}{1-X} =! + X + X^2 + X^3 + \dots$$

with

$$X = \sum_{n>0} \left(-\frac{a_n}{a_0} \right) x^n.$$

For each j, the coefficient of x^n in

$$\left(\sum_{n>0} \left(-\frac{a_n}{a_0} \right) x^n \right)^j$$

is

$$\sum \binom{j}{t_1, t_2, \ldots t_n} \left(-\frac{a_1}{a_0} \right)^{t_1} \cdots \left(-\frac{a_n}{a_0} \right)^{t_n},$$

where summation is extended for all partitions of n, and

$$\binom{j}{t_1, t_2, \ldots t_n}$$

is the associated multinomial coefficient. In conclusion:

Theorem: Trudy's Formula

$$\begin{vmatrix} a_1 & a_0 & 0 & \ldots & 0 \\ a_2 & a_1 & a_0 & \ldots & 0 \\ a_3 & a_2 & a_1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ldots & 0 \\ a_{n-1} & a_{n-2} & \ldots & a_1 & a_0 \\ a_n & a_{n-1} & a_{n-2} & \ldots & a_1 \end{vmatrix} = \sum \binom{t_1 + t_2 + \cdots + t_n}{t_1, t_2, \ldots t_n} (-a_0)^{n - t_1 - t_2 - \cdots - t_n} a_1^{t_1} a_2^{t_2} \ldots a_n^{t_n}$$

(Reference: Merca, Special Matrices, also Muir)

Kaluza's Sign Condition

Let

$$P := a_0 + a_1 x + a_2 x^2 + \ldots$$

be a formal power series (analytic function) over the field of real numbers such that

$$a_n > 0,\ n = 0, 1, 2, \ldots$$

and

$$a_{n+1}a_{n-1} - a_n^2 > 0,\ n = 1, 2, \ldots$$

If

$$P^{-1} := b_0 + a_1 x + b_2 x^2 + \ldots$$

then show that

$$b_n < 0,\ n = 1, 2, \ldots.$$

Preliminaries

Complete solution by Carlitz, American Mathematical Monthly,66(5),Problem 4803.

Quick background search reveals the following: A sequence of non-negative numbers

$$a_0, a_1, a_2, \ldots$$

is called *convex* if

$$\frac{a_{n+1} + a_{n-1}}{2} \geq a_n$$

and *log-convex* if

$$\sqrt{a_{n+1}a_{n-1}} \geq a_n.$$

By arithmetic-geometric inequality

$$\frac{a_{n+1} + a_{n-1}}{2} \geq \sqrt{a_{n+1}a_{n-1}} \geq a_n,$$

or log-convexity implies convexity. Moreover

$$a_{n+1} * a_{n-1} \geq a_n^2 \ \Rightarrow \frac{a_{n+1}}{a_n} \geq \frac{a_n}{a_{n-1}}$$

$$a_n * a_{n-2} \geq a_{n-1}^2 \;\Rightarrow \frac{a_n}{a_{n-1}} \geq \frac{a_{n-1}}{a_{n-2}}$$

hence

$$\frac{a_{n+1}}{a_n} \geq \frac{a_{n-1}}{a_{n-2}} \cdots \geq \frac{a_{n-r+1}}{a_{n-r}}.$$

A sequence of positive numbers is log-convex if and only if the sequence of $\left\{ \frac{a_{n+1}}{a_n} \right\}$ is increasing. Further, *log-concavity* is defined by

$$\sqrt{a_{n+1}a_{n-1}} \leq a_n.$$

It is known that the log-convexity of a positive sequence is equivalent to the log-concavity of the reciprocal series. In the problem under consideration we claim that strict log-convexity,

$$\sqrt{a_{n+1}a_{n-1}} > a_n, \quad n = 1, 2 \ldots.$$

implies sign-regularity,

$$b_n < 0, \; n = 1, 2 \ldots.$$

Discussion:

Without loss of generality, set a_0=1. It is well known that the canonical convolution of formal power series P with its inverse

$$PP^{-1} = 1,$$

can be represented by a product of an infinite triangular matrix (or array) by an infinite column vector:

$$\begin{bmatrix} a_0 & & & & & \\ a_1 & a_0 & & & & \\ a_2 & a_1 & a_0 & & & \\ a_3 & a_2 & a_1 & a_0 & & \\ a_4 & a_3 & a_2 & a_1 & a_0 & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \\ \cdots \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ \cdots \end{bmatrix}$$

From this system of linear equations we obtain recursively, by Cramer's rule,

$$b_0 = \frac{1}{a_0}, \quad b_0 = 1.$$

$$b_1 = -\frac{a_1}{a_0^2}, \quad b_1 = -a_1 < 0, \surd$$

$$b_2 = \frac{1}{a_0^2}\begin{vmatrix} a_1 & a_2 \\ a_0 & a_1 \end{vmatrix} = \frac{1}{a_0^3}(a_1 a_1 - a_2 a_0) < 0, \surd$$

by construction. However, we cannot tell by inspection that

$$b_3 = -\frac{1}{a_0^4}\begin{vmatrix} a_1 & a_2 & a_3 \\ a_0 & a_1 & a_2 \\ 0 & a_0 & a_1 \end{vmatrix}$$

is less than zero. So we close this line of inquiry with the meager result : $b_1 < 0,\ b_2 < 0$.

Can we determine the sign of b_n if we know

$$b_0 > 0,\ b_k < 0, k = 1, 2, \ldots (n-1)?$$

We shall examine this by a numerical example. Set

$$a_k = \frac{1}{1+k},\ k = 0, 1, 2, \ldots$$

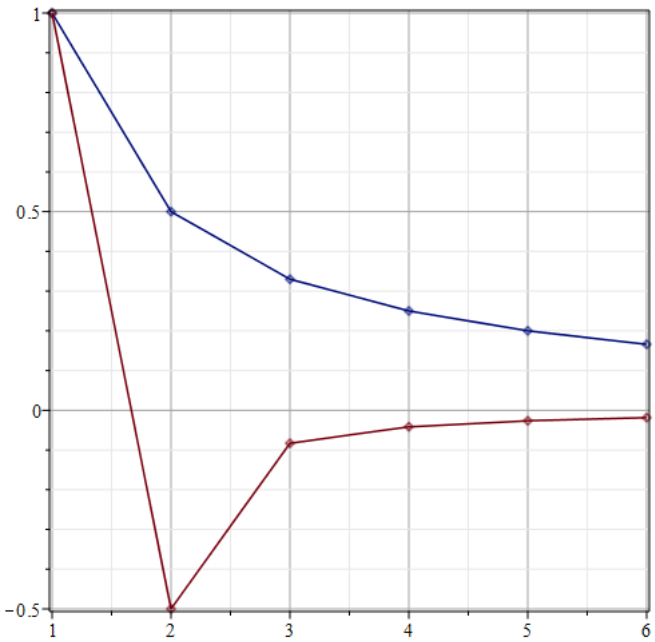
The log-convexity condition is satisfied

$$a_{k+1}a_{k-1} - a_k^2 > 0,\ k = 1, 2, \ldots$$

and the data is as follows:

k	a_k	b_k
0	1	$+1$
1	2^{-1}	-0.5
2	3^{-1}	$-8.33E-2$
3	4^{-1}	$-4.16E-2$
4	5^{-1}	$-2.64E-2$
5	6^{-1}	$-1.86E-2$

The simple line chart of a_k and b_k is



where $a_k,\ k = 0, 1, \ldots$ is a log-convex (hence convex) strictly monotone decreasing sequence and $b_k,\ k = 1, 2 \ldots$ is a strictly monotone increasing concave sequence, which has the x-axis its asymptote. This of course, is not a proof, it is just a simple observation of some numerical data. There is no new information here. Clearly, it would e desirable to get rid of b_0 , $b_0 > 0$ while $b_k < 0,\ k =, 2 \ldots n$.

Let us try writing out equations for b_n and b_{n+1}

$$\begin{aligned} a_0 b_n &= -a_1 b_{n-1} - a_2 b_{n-2} - \cdots - a_n b_0 \\ a_0 b_{n+1} &= -a_1 b_n - a_2 b_{n-1} - \cdots - a_{n+1} b_0. \end{aligned}$$

or

$$0 = -a_0 b_n - a_1 b_{n-1} - a_2 b_{n-2} - \cdots - a_n b_0 \tag{21.2}$$

$$a_0 b_{n+1} = -a_1 b_n - a_2 b_{n-1} - a_3 b_{n-2} - \cdots - a_{n+1} b_0. \tag{21.3}$$

Next, eliminate b_0 from the equations. Multiply the first equation by $-a_{n+1}$ and the second equation by a_n and add them:

$$\begin{aligned} 0 &= a_{n+1} \left(a_0 b_n + a_1 b_{n-1} + a_2 b_{n-2} + \ldots \right) + a_{n+1} a_n b_0 \\ a_n a_0 b_{n+1} &= -a_n \left(a_1 b_n + a_2 b_{n-1} + a_3 b_{n-2} + \ldots \right) - a_n a_{n+1} b_0. \end{aligned}$$

the resulting in

$$\begin{aligned} a_n a_0 b_{n+1} &= (a_{n+1}a_0 - a_n a_1)\, b_n + (a_{n+1}a_1 - a_n a_2)\, b_{n-1} + \ldots \\ &+ (a_{n+1}a_{n-1} - a_n a_n)\, b_1. \end{aligned}$$

Coefficient of b_n:

$$a_{n+1}a_0 - a_n a_1 > 0$$

$$a_{n+1}a_0 > a_n a_1$$

$$\frac{a_{n+1}}{a_n} > \frac{a_1}{a_0}\surd$$

Coefficient of b_{n-1}:

$$a_{n+1}a_1 - a_n a_2 > 0$$

$$a_{n+1}a_1 > a_n a_2\surd$$

$$\vdots$$

Coefficient of b_1:

$$a_{n+1}a_{n-1} - a_n a_n > 0\surd$$

Since $b_k < 0,\ k = 1, 2, n$ and $a_k > 0,\ k = 0, 1, 2, n$ it follows that b_{n+1} is negative. (In other words, sign regularity is proven .)

where $(-a_1 b_1)$ is positive and $(-a_2 b_0)$ is negative. We can observe further that in

$$a_0 b_3 = -a_1 b_2 - a_2 b_1 - a_3 b_0$$

the terms $(-a_1 b_2 - a_2 b_1)$ have a positive contribution, whereas $(-a_3 b_0)$ is positive. Moreover, if we reduce the third equation to zero we notice that $b_2,\ b_1,\ b_0$ appears on the right hand sides

$$0 = a_0 b_2 - a_1 b_1 - a_2 b_0 \tag{21.4}$$

$$a_0 b_3 = -a_1 b_2 - a_2 b_1 - a_3 b_0 \tag{21.5}$$

Let us try elimination:multiply the first equation by $-a_3$ and the second equation by a_2 and add them:

$$\begin{aligned} 0 &= a_3a_0b_2 + a_3a_1b_1 + a_3a_2b_0 \\ a_2a_0b_3 &= -a_2a_1b_2 - a_2a_2b_1 - a_2a_3b_0 \\ a_2a_0b_3 &= \end{aligned}$$

Next, let us rewrite the last two equations by inserting $a_0 = b_0 = 1$

$$0 = b_n - a_1b_{n-1} - a_2b_{n-2} - \cdots - a_n$$

$$b_{n+1} = -a_1b_n - a_2b_{n-1} - \cdots - a_{n+1}.$$

$$0 = b_n - a_1b_{n-1} - a_2b_{n-2} - \cdots - a_n \tag{21.6}$$

$$b_{n+1} = -a_1b_n - a_2b_{n-1} - \cdots - a_{n+1}. \tag{21.7}$$

Therefore we rewrite the system of equations in the original format by Kaluza.:

$$\begin{bmatrix} 1 & & & & & \\ a_1 & 1 & & & & \\ a_2 & a_1 & 1 & & & \\ a_3 & a_2 & a_1 & 1 & & \\ a_4 & a_3 & a_2 & a_1 & 1 & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix} \begin{bmatrix} 1 \\ -c_1 \\ -c_2 \\ -c_3 \\ -c_4 \\ \cdots \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ \cdots \end{bmatrix}$$

with

$$a_0 = b_0 = 1, \quad b_n = -c_n \ (n \geq 1).$$

Line-by-line

$$\begin{aligned} 1 * 1 &= 1 \\ a_1 - c_1 &= 0 \\ a_2 - a_1c_1 - c_2 &= 0 \end{aligned}$$

Hence

$$\frac{1}{\sum_{n=0}^{\infty} a_n x^n} = 1 - \sum_{n=1}^{\infty} c_n x^n.$$

$$0 = a_n - \sum_{r=1}^{n} c_r a_{n-r}, \quad (n \geq 1) \tag{21.8}$$

(Revised and updated 1x.)